

# Z6 Narzędzia do wydobywania i wizualizacji informacji z tekstu

Z6.1. Korpus treningowo-testowy znakowany w zakresie atrybutów wyznaczników sytuacji i relacji między wyznacznikami sytuacji

Okres sprawozdawczy: M18

Politechnika Wrocławska

*Autorzy:*

Marcin Oleksy

Jan Kocoń

Michał Marcińczuk

# 1. Wstęp

Celem zadania było opracowanie korpusu treningowo-testowego znakowanego w zakresie atrybutów wyznaczników sytuacji i relacji między wyznacznikami sytuacji. Będzie to służyć opracowaniu narzędzia, które będzie rozpoznawać strukturę zdarzeń ze szczególnym uwzględnieniem określenia faktycznego wystąpienia sytuacji.

## 2. Realizacja

Pierwszym etapem realizacji zadania było opracowanie opartych na specyfikacji TimeML wytycznych ręcznego oznaczania relacji między sytuacjami oraz ich atrybutów. Procesowi temu służyła anotacja tekstów służąca wychwyceniu niespójności i punktów dyskusyjnych. Znakowaniu podlegał korpus tekstów KPWr anotowanych uprzednio w zakresie wyznaczników sytuacji, a do tworzenia zasobu wykorzystano system Inforex. Przeprowadzono kilka iteracji znakowania, mierząc zgodność między anotatorami. Współczynnik osiągnięty w trakcie iteracji o najniższej zgodności wynosił 0.75, najlepsza zaś pod tym względem iteracja wiązała się z osiągnięciem zgodności 0.91. Zgodność między anotatorami mierzona dla całego zbioru wynosi 0.85. Anotacja przebiegała w systemie 2+1, a do finalnego zbioru trafiły relacje wskazane przez obu anotatorów lub pozytywnie zweryfikowane przez superanotatora.

Korpus został wzbogacony również ręcznie przypisanymi atrybutami wyznaczników sytuacji. W procesie anotacji wykorzystano zarówno techniczną możliwość przypisywania atrybutów anotacjom, jak i relacje między anotacjami, które pozwoliły na powiązanie z tekstowymi wykładnikami atrybutu *modality*.

## 3. Rezultat

Proces znakowania korpusu KPWr zakończył się opracowaniem zasobu składającego się z 351 dokumentów anotowanych w zakresie relacji między wyznacznikami sytuacji (por. Tab 1). Ponadto relacje tego typu zostały wprowadzone w wyselekcjonowanym zestawie dokumentów pochodzących z korpusu Polish Spatial Texts stanowiącego korpus treningowo-testowy znakowany dynamicznymi wyrażeniami przestrzennymi.

W opublikowanym zbiorze dokumentów 3044 wyznaczniki sytuacji mają przypisane atrybuty (szczegółowe statystyki zostały przedstawione w Tab. 2). Każdy z nich ma przypisane wartości dwóch atrybutów (*polarity* i *generality*), tam zaś, gdzie w tekście pojawiło się odpowiednie wyrażenie, wyznacznikowi przypisywano wartość atrybutu *modality*.

nazwa relacji	opis	liczba instancji
ALINK	relacje pomiędzy czasownikami należącymi do klasy ASPECTUAL a ich argumentami zdarzeniowymi	304
SLINK	relacje pomiędzy dwoma zdarzeniami, z których jedno ma charakter modalny, faktywny bądź poświadczający	1808
razem		

Tab. 1 Finalne relacje między wyznacznikami sytuacji

nazwa atrybutu	liczba instancji
polarity (biegunowość)	3044
generality (ogólność)	3044
modality (modalność)	181
razem	

Tab. 2 Liczba wyznaczników sytuacji z przypisanymi atrybutami